Recommending Citations for Academic Papers

Trevor Strohman
strohman@cs.umass.edu

W. Bruce Croft
croft@cs.umass.edu

David Jensen
jensen@cs.umass.edu

Department of Computer Science
University of Massachusetts
Amherst, MA 01003

ABSTRACT

We approach the problem of academic literature search by considering an unpublished manuscript as a query to a search system. We use the text of previous literature as well as the citation graph that connects it to find relevant related material. We evaluate our technique with manual and automatic evaluation methods, and find an order of magnitude improvement in mean average precision as compared to a text similarity baseline.

Categories and Subject Descriptors: H3.3 Information Storage and Retrieval: Information Search and Retrieval

General Terms: Design, Experimentation

Keywords: Bibliometrics

1. INTRODUCTION

Most current literature search systems concentrate on short queries that are unlikely to describe fine details of the user's true information need. In this work, we instead suppose that the user is able to provide the system with a very long query; we assume that the user has already written a few pages about the topic, and is able to submit this document to the search system as the query. We conjecture that this additional information can improve the effectiveness of the ranked list of documents. Instead of assuming that the user wants documents that are topically similar to the query, we assume the user wants documents that the query document might cite. This is particularly challenging because the concept of relevance is much stricter than in ad hoc retrieval; most papers could cite hundreds of topically similar papers, but contain just a few highly relevant citations.

We have built a system to explore this citation recommendation problem. In the process, we have found that simple text similarity computation is not enough for this task. We show that it is necessary to use graph-based features in the retrieval process to achieve high quality retrieval results, although many seemingly useful features offer little benefit.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.

Publication Year	The year the document was published (normalized by subtracting 1950)
Text Similarity	The similarity of the text of this candidate with the query, as measured by the multinomial diffusion kernel [1]
Co-citation Coupling	The fraction of documents that cite this candidate that also cite documents in the base set
Same Author	Binary feature; true if this document is written by the same author that wrote the query
Katz	The Katz graph distance measure [2]: $\sum_i \beta^i N_i$, where N_i is the number of unique paths of length i between the two nodes, and β is a decay parameter between 0 and 1.
Citation Count	Number of citations of this document from all documents in the corpus

Table 1: Features used in our experimental model

2. MODEL

Our system uses a two stage process to find a set of documents to rank. In the first step, the system searches a collection of over a million papers, and returns the top 100 most similar papers to the query document as the set R. In the second step, all papers cited by any paper in R are added to R. In general, this process concludes with a set R that contains 1000 to 3000 documents. Initial experimentation with real academic papers suggested that over 90% of papers that researchers actually cite would be in R at this point. Expanding R with a third step (again adding all papers that are cited by some paper in R) did not appear improve recall.

We then rank the documents in R by the features shown in Table 1. Neither text-based nor citation-based features performed well in isolation. Text-based features are good for finding some similar related work. However, text features are not as good at finding conceptually related work that uses different vocabulary. Textual features are also poor at establishing authority of documents. Citation features are useful for these things, but may do a poor job at coverage (since recent documents may have no citations).

We use coordinate ascent to find feature weights for our model. The features are combined in a weighted linear model to provide a document score, which is used to rank the documents in R.

3. EVALUATION

To evaluate this system, we treated published research papers as queries. These papers were drawn from an early copy

		Full			Truncated		
		Mean	Interval		Mean	Interval	
Baseline	Text Similarity	0.0079	0.0055	0.0103	0.0079	0.0055	0.0103
Experimental	All Features	0.1016	0.0781	0.1251	0.0940	0.0727	0.1153
	No Text	0.0675	0.0539	0.0811	0.0612	0.0469	0.0754
	No Author	0.0983	0.0747	0.1219	0.0917	0.0701	0.1132
	No Katz	0.0335	0.0256	0.0414	0.0257	0.0194	0.0320
	No Cite Count	0.1005	0.0771	0.1238	0.0931	0.0718	0.1144
	No Date	0.1052	0.0834	0.1269	0.0979	0.0784	0.1174
	No Title	0.1016	0.0781	0.1251	0.0940	0.0727	0.1153

Table 3: Results of 10-fold cross validation experiments on a 1000 query set. Results are reported using the mean average precision metric. Full results represent mean average precision over the entire retrieved set, while the truncated results reflect mean average precision computed over the first hundred retrieved documents. Confidence intervals are based on the t distribution over all 10-folds. All experimental models significantly outperform text similarity (Wilcoxon, $p = 0.01$). All experimental models with the Katz measure significantly outperform the "No Katz" method (Wilcoxon, $p = 0.01$)

Total paper entries	964,977
Papers with text	105,601
Total number of citations (X cites Y)	1.46 million
Total number of cited papers	675,372

Table 2: Statistics from the Rexa collection used in our experiments

of the Rexa[1] database (Table 2). Note that while there are almost a million entries in this collection, only about 10% of them contain the full text of the paper. We performed a small manual evaluation of search result quality, but space restrictions keep us from reporting those results here.[2]. To evaluate the system without manual intervention, we considered the references list from the query paper as the relevant citations, then evaluated our retrieval system on its ability to find the references in this list. We chose 1000 documents from the Rexa collection to use as sample queries. In order to have the best possible generalization to full text collections, we chose query documents where a large percentage of their citations were full-text Rexa entries.

We used a text similarity baseline, which is the first stage of our experimental algorithm, with no additional features. Since other models may return more than 100 documents, we also perform a truncated evaluation for each model, where only the top 100 documents are considered. The truncated column allows a fair comparison between the text similarity baseline and the other models.

In order to assess the usefulness of particular features, we performed experiments that removed each feature from the model in isolation. We expect that if a feature is very useful, the retrieval effectiveness of the system will drop dramatically when a feature is removed; if it is not useful, we expect effectiveness to stay the same. Note that we did not re-train the model for these tests; we only set the weight of the removed feature to zero.

3.1 Results

The results of our experiments are shown in Table 3. Our experimental results show the effectiveness of our system in various modes against a text similarity baseline. The confidence intervals come from the t distribution. We also performed the distribution-free Wilcoxon signed rank test ($p < 0.01$) for significance. From this, we find that all experimental models significantly outperform the text similarity baseline. Also, we find that the "No Katz" experimental model is significantly outperformed by all other experimental models ($p < 0.01$). The truncated "No Text" is significantly outperformed by all models with both the Katz feature and Text ($p < 0.05$), although we can conclude nothing about the "No Text" non-truncated model.

Surprisingly, text similarity alone is a poor way to succeed at this task. The baseline results are very low by information retrieval standards, but to succeed in this task, the system must find not just related work, but the most influential and highest quality work. The citation features play a major role in finding these high quality documents.

A second surprise is how little many of the features we used matter in the final ranking of documents. The author, citation count, publication date and title text features add little to nothing to the effectiveness of the system. This is not to say that these features are not correlated with relevance, but they are dominated by the full text and Katz features.

The Katz measure is crucial to the performance of our model. Without the Katz feature, model performance drops by over half. One way to interpret this result is that the Katz measure is closest in capturing what scientists actually cite.

[1]http://www.rexa.info

[2]Full details in the technical report version of this paper.

4. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant #CNS-0454018, in part by The CIA, the NSA and NSF under NSF grant #IIS-0326249, and in part by ARDA and NSF grant #CCF-0205575. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

5. REFERENCES

[1] J. Lafferty and G. Lebanon. Diffusion kernels on statistical manifolds. *JMLR*, 6:129–163, 2005.

[2] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM 2003*, pages 556–559, 2003.